DOCUMENT RESUME

ED 356 238 TM 019 647

AUTHOR Beaton, Albert E.
TITLE Considerations for National Examinations. A Policy Issue Perspective.
INSTITUTION Educational Testing Service, Princeton, NJ. Policy Information Center.
PUB DATE 92
NOTE 19p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Standards; *Accountability; Educational Assessment; *Educational Objectives; Elementary Secondary Education; Evaluation Methods; National Competency Tests; *National Programs; Position Papers; Student Evaluation; *Systems Development; Test Construction; *Testing Programs; Test Use
IDENTIFIERS Scholastic Aptitude Test; *Standard Setting

ABSTRACT

Neither teacher-made tests nor the Scholastic Aptitude Test are appropriate for measuring the attainment of national education goals. Most practitioners involved in discussions of national high-stakes examinations argue that new tests are needed. A new system that preserves state and local autonomy in education as it measures the attainment of national and world-class standards has been proposed. Several issues in developing a national testing system are discussed. The first question is what the test would cover. It makes sense to develop alternative cluster tests for different curricula and to set appropriate standards for each cluster. Deciding who will take the test and how to compare results is a second set of issues to be resolved. Setting performance standards is another question with important consequences. A different set of problems is posed in using a test for accountability. Many specific questions must be addressed in the implementation of a national testing system. A test intended for many purposes will be unlikely to serve any purpose very well. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Coley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

Copyright © 1992 by Educational Testing Service

Considerations

For

National Examinations

by

Albert E. Beaton

Policy Information Center
Educational Testing Service
Princeton, New Jersey 08541

The views expressed in this report are those of the author and
not necessarily of the ETS Policy Information Center.

Table of Contents

	Page
Preface	i
Acknowledgments	ii
Introduction	1
Educational Tests: What Can They Do?	2
Issues in Developing a National Testing System	4
Who Will Take the Tests?	6
Performance Standards: Who Decides?	8
Accountability: Questions Requiring Answers	10
Getting a Grip on the Specifics	10
References	12

Preface

We asked Professor Albert E. Beaton to draw on his long experience in statistics and education measurement to reflect on considerations in constructing a nationwide system of individual examinations and assessments. He recently joined the Center for the Study of Testing, Evaluation and Public Policy at Boston College. Prior to that he was Director of Design, Research, and Data Analysis for the National Assessment of Educational Progress at Educational Testing Service.

Professor Beaton took on the difficult task of presenting the statistical and psychometric considerations in constructing such a system in terms and language that can be understood by a nontechnical audience. As is so often the case, the judgments must be made by policy and education officials who are not masters of the technical underpinnings of educational measurement, but who need to have enough comprehension of these matters to make informed choices.

Paul E. Barton
Director
Policy Information Center
January 1992

Acknowledgments

A number of eminent educational researchers have reviewed early drafts of this manuscript and made helpful comments to the author. These researchers were: Paul Barton, Nancy Cole, John Fremer, Robert Linn, George Madaus, Robert Mislevy, and Ina Mullis. The author is grateful for these comments, although he accepts full responsibility for the final content. He is also indebted to the editor, Joanne Pfleiderer, of the Educational Testing Service. Carla Cooper of the Policy Information Center provided the desktop publishing services.

Albert E. Beaton

Introduction

The call for educational reform in the United States has become a roar. Educational policymakers are deeply dissatisfied with the performance of students in American schools. Business leaders complain that recent high school graduates do not have the skills needed to succeed in the work force, which threatens the nation's international position. University faculties complain that, because many first-year students are not prepared for the rigors of a university education, remedial programs must be provided.

Many have judged the proficiency levels of American students, documented by the National Assessment of Educational Progress (NAEP), to be wanting. It is widely believed that schools are not meeting the needs of our increasingly technological, democratic society.

Recently, the reform movement has targeted high-stakes testing as a way of:

- Driving school curricula.
- Motivating students to aspire to new, higher standards.
- Motivating teachers to assure that students reach their goals.

The high-stakes testing movement seeks to establish challenging tests in each subject area, with high standards set for student performance.

The high-stakes testing movement seeks to establish challenging tests in each subject area, with high standards set for student performance. According to this approach, the test results would be used for a number of important decisions about students, such as promotion to the next grade level, graduation, honors, employment, and university admission. Because the tests would have such high stakes, students would be motivated to work harder to achieve the rewards associated with high performance. Teachers would be pressured to assure that their students performed well and would teach to the subject matter and levels of performance required to pass the tests. Ultimately, the tests would shape instruction and commonly-held views of acceptable levels of performance. Teachers would have substantial freedom in teaching methods, as long as the specified goals were attained.

This test-based reform movement relies heavily on the assumption that a national testing system can be designed, administered, and interpreted in ways that would improve the educational process and student performance.

This test-based reform movement relies heavily on the assumption that a national testing system can be designed, administered, and interpreted in ways that would improve the educational process and student performance. In building this system, the technology and limitations of testing will come into play. Cost and feasibility will force trade-offs that will have to be addressed. Proposals for a national testing system are changing as the interplay of educational policy and technology is examined. This paper explores some of the basic issues in building and using high-stakes tests in educational reform and recommends some ways to proceed.

Educational Tests: What Can They Do?

To measure what students know and can do, an educational test asks a student to complete a task or series of tasks under standard conditions. These standard conditions ensure that all test takers have an equal opportunity to do well and that scores are comparable. The tasks are often direct questions requiring written, open-ended, or multiple-choice responses. Student activities, such as preparing notebooks, collecting portfolios, or performing laboratory experiments, may also be considered tests. The tasks in a test typically represent only a small sample of what has been taught. Through their responses to the tasks, we generalize what students have learned more broadly and to tasks not included on the test. Standards are applied to establish levels of performance.

Tests vary markedly in form and purpose, so it may be instructive to compare two quite different types:

- *Classroom tests.* Most practicing teachers prepare, administer, score, and grade classroom tests on a regular basis. Test content is closely tied to what has been recently taught, and tests are given only to students who have covered the material involved. The teacher may make up the questions or use those published in teachers' editions of textbooks. Questions may require simple recall, such as names or dates, or higher order thinking, such as an original proof in geometry. Teachers grade the answers using their own scoring keys and performance standards. Since scoring is done by the teacher, turnaround is as fast as the teacher chooses.

The information generated by classroom tests can have many uses, including providing general information about what the teacher taught well, as well as what the class learned well. Results can also provide diagnostic information about individual learning problems, and information for student progress reports to parents and for preparing grades that are part of permanent school records. In high school, a student's grade point average summarizes performance from test scores and other considerations.

- *College admission tests, such as the Scholastic Aptitude Test (SAT).* The SAT is very different from classroom tests in form, substance, and purpose. As a national test used in the college admission process, the SAT is taken largely by college-bound students, not the general high school population. Because college-bound students in the United States do not follow a common curriculum, the test cannot cover specific content without giving an advantage to some students. A case in point: Before the use of the SAT became widespread, some college entrance requirements included proficiency in specific subjects, such as Latin. The result? Students in public schools that did not teach Latin were essentially eliminated from consideration for college admission. Because colleges could not enforce a common national curriculum, the largely curriculum-free SAT became useful as a predictor of applicants' success in college. The SAT succeeded by including questions based on fairly basic to higher-order reasoning skills that could be answered without the detailed knowledge that comes from studying specific high school subjects.

  To be as fair and impartial as possible, the SAT is administered under strictly controlled, standardized conditions. There is no single standard for acceptable performance. Colleges use SAT results in conjunction with other information, such as high school grade point average, to set their own standards for admission.

Neither teacher-made tests nor the SAT are appropriate for measuring the attainment of the national goals. Although content relevant, teacher-made tests are not comparable across the country and across curricula. The SAT is comparable across

curricula, but does not directly measure specific student learning and proficiency in school subjects. A different test is needed to accomplish the purposes of education reform. As explained later, there is no such thing as an all-purpose test.

...there is no such thing as an all-purpose test.

Issues in Developing a National Testing System

Most of those involved in discussions of r tional, high-stakes examinations argue that new tests, falli somewhere in between and substantially different from the two already quite different tests described above, are needed. A new examination system that preserves state and local autonomy in education but also measures the attainment of national and world-class standards has been proposed. This new system may involve "cluster" testing, allowing states or groups of states--and perhaps local districts--to develop exams coordinated with their common curriculum.

The proposed system would be voluntary, at least in the sense that the states would choose a particular test and administer it to their students. School systems could use the tests for different purposes. Each student would be graded individually, and rewards or sanctions would result from test scores. Performance would be judged by standards that were high enough, and sanctions that were strong enough, to assure that students were motivated to learn and teachers to teach.

Performance would be judged by standards that were high enough, and sanctions that were strong enough, to assure that students were motivated to learn and teachers to teach.

To make the standards the same for tests administered by different clusters of states, a national "anchor" test has been suggested. This test would be administered along with or as part of the cluster tests and used as a basis for comparing them. The educational system as a whole could then be held accountable for meeting national standards.

Developing a national testing system raises many questions. Some of them can be answered by our experiences with existing tests, such as teacher-made tests and a national test like the SAT.

The first question is, what would the test cover? As we have seen, teacher-made tests assess recently taught material, while the SAT evaluates general reasoning skills. A national curriculum at the elementary level, covering reading, writing, and arithmetic, is plausible, as are reasonably fair tests to measure these accomplishments, although there is a problem for national testing if these subjects are taught in a different order in different places. Although school systems may use different content to teach these skills, they can be tested and

compared if the test is not so dependent on content that it unfairly advantages or disadvantages anyone. Different tests tailored to specific curricula lead to problems in equating, calibrating, and projection that will be discussed later. Although developing common examinations at the elementary level seems feasible, vast differences in student performance still exist. James Coleman demonstrated in 1966 that these differences are present even at the beginning of first grade.

Differences in content and levels of performance become more striking as students progress through school. The intermediate school curriculum is more differentiated than the elementary. Curriculum differentiation at the secondary school level is quite large. Although all subject areas display some differentiation at this level, the mathematics curriculum is an extreme example.

- Calculus in advanced college preparation programs.
- Bookkeeping in business programs.
- Specialized mathematics in vocational programs.

A single mathematics examination for all secondary school students could not reasonably cover all the material taught in the different curricula. And attempting to build such a test would not be wise. A relevant testing instrument would measure the calculus knowledge of students who had taken calculus, and would not be given to students who had not studied the subject. Unless all secondary students were to take a common mathematics curriculum, a national test would be so general that no particular subject area would be covered adequately. In high school mathematics, a test this broad would reduce, not increase, the incentive to study calculus, and could lower the accomplishments of the most advanced students.

Under the present system of different curricula for different students, it makes sense to develop alternative cluster tests for different curricula, and set appropriate standards for each cluster. Along with these tests, an effort could be made to encourage more students to attempt the more challenging curricula. All students would not, however, be trying to meet the same standards.

Who Will Take the Tests?

If states or other jurisdictions are to be compared on test performance, then the populations of students must be reasonably similar for the comparisons to be fair and useful as accountability tools.

Administering several different tests in a subject area brings up the problem of who will take the tests. If states or other jurisdictions are to be compared on test performance, then the populations of students must be reasonably similar for the comparisons to be fair and useful as accountability tools.

Teacher-made tests do not address this problem, since there is no formal attempt to compare work in one school to that in another. Even within the same school, some teachers are perceived as "easy" and others as "hard" markers. Although it measures reasoning skills that are comparable across test takers and has merit as an indicator of the competence of college-bound students, the SAT is not suitable as a general national educational performance indicator either, since SAT takers consist largely of college-bound students and, more specifically, students who are applying to certain colleges. No useful way has been found to estimate general student performance through SAT results. NAEP has been able to compare the mathematics proficiency of eighth graders across states, through its general mathematics assessments administered to random samples of eighth graders in participating states.

If pupils studying different curricula at varying levels of difficulty were administered different tests, it would be difficult to compare the results. The basic problem is a lack of fairness and meaning in the results. For example, suppose only the top 20 percent of students in State A took advanced mathematics, while 75 percent of the students in State B did so, particularly because that state aggressively encouraged this curriculum. Comparing the results of these two states would be analogous to comparing the average score of the general American high school population with the average score of elite students in another country; it just isn't fair or meaningful. State B would be very likely to look worse than State A on a test of advanced mathematics given to students taking advanced mathematics. In terms of this comparison, State B would look better if it encouraged poorer performing students to enter less challenging courses so that they might score higher.

Clearly, evaluating group performance is a complex matter that must consider not only student performance on a particular test but also the percentage of students taking the test. Even comparing tests of generally equal difficulty is problematic when the tests measure different curricula. Consider, for example, two different curricula for advanced mathematics; one requires trigonometry in the twelfth grade,

and the other requires solid geometry. Different examinations could be developed to measure student performance in trigonometry and solid geometry. But how would we know which students were more advanced in mathematics? It's like asking whether baseball great Babe Ruth or football giant Red Grange was the better athlete.

Test developers can use several techniques to compute types of correspondence of two different tests even though they do not measure the same proficiency.

Test developers can use several techniques to compute types of correspondence of two different tests even though they do not measure the same proficiency. In 1991 Martha Stocking and Robert Mislevy described three different types of test correspondence:

- *Equating*, the strongest form of correspondence, is appropriate when the two tests are very similar in content and design.

- *Calibration* is suited to two tests measuring the same proficiency, but doing so in different ways. Stocking and Mislevy show that some comparisons of calibrated results are appropriate and others are not.

- *Projection* is the weakest form of comparison. It can be used to estimate how a student who performed at a specific level on one test would do on the other, even though the tests measure different things. It is based on specific assumptions about the relationships between the tests.

Equating results are truly satisfactory only in a few cases. Equating works well for the SAT because each form of the test measures the same ability and is developed according to the same specifications. Equating ensures that scores on different forms of the SAT, which are administered at different times of year, are strictly comparable.

If two tests are not measuring the same ability, then comparing scores becomes questionable. For example, in a 1977 study of the decline in SAT scores, two apparently similar reading tests with different numbers of items were calibrated, and it was found that the longer test was substantially more accurate in identifying the better and the poorer students than was the shorter test.

Projection can be used for different tests, and so we may guess that a student who did well on one test would also do well on the other, if he or she had studied the subject, but with this weak method the inference is highly speculative.

An anchor test can provide information about the equivalence of two different tests by measuring more general skills or something that is common to both tests. Information from the anchor can be used to bridge between the two tests, at least in a rough way.

An anchor test can provide information about the equivalence of two different tests by measuring more general skills or something that is common to both tests. Information from the anchor can be used to bridge between the two tests, at least in a rough way. For example, if trigonometry students did better than solid geometry students on a general mathematics anchor test, we could assume that trigonometry was offered to or selected by the more mathematically proficient students--or did more to improve general mathematical ability. We could then judge a top score in trigonometry to be more outstanding than a similar score in solid geometry.

However, giving an anchor test of general mathematics skills to both student groups would be like comparing Ruth and Grange on how fast they could run 90 feet, 100 yards, and how many pounds they could bench press. The standard of comparison can be used to gauge roughly general ability but is a pale reflection of specific skills. A value judgment would have to be made on whether the rough comparison is good enough, and that judgment would be informed by the test results. Problems in equating the contents and levels of different tests make it desirable to judge student performance by the standards associated with a particular test. Without a common curriculum, comparing across curricula is speculative at best.

Problems in equating the contents and levels of different tests make it desirable to judge student performance by the standards associated with a particular test. Without a common curriculum, comparing across curricula is speculative at best.

Robert Linn suggested in 1991 an alternative way of attaining comparable standards with differing curricula. In his model, one cluster of states or districts decides on one curriculum and test, while another cluster decides on a different curriculum and test. Each cluster grades a sample of the other cluster's test results, exchanging information about performance and coming to an understanding of the comparability of results. A negotiating process is involved in deciding on appropriate levels of performance. While this process could be used to compare many different clusters of curricula and their tests, its adequacy has yet to be explored.

Performance Standards: Who Decides?

Setting performance standards is also tricky and has important consequences. With teacher-made tests, teachers set standards informally, based on what can be reasonably expected of students. College admission examinations leave standard setting to colleges, which decide the levels of performance necessary for students to succeed at that college. There are many ways to set standards. For example, university standards in Great Britain are rigorously maintained by an external examiner system. Before any course is taught in a British

university, a professor at another university reviews the syllabus as well as the final examination questions and acceptable answers. This external examiner, who must assure that the content and level of the course is appropriate, also reviews final examinations and has the right to change grades he or she deems inappropriate. This process allows the university system to maintain high, common standards.

It is hard to envision this system being applied in the United States--especially difficult if it implies that a Yale professor can decide what is taught at Harvard. But the system evokes a side effect of maintaining high standards: Some students will fail. In the British system, those who fail simply drop out of the university system, since failing in one university is tantamount to failing at all. In the United States, we want every student to have the opportunity to attend college. Because universities have different standards, failure to be admitted at one university does not imply that a student cannot attend and succeed at another. The same is true at the lower grade levels. We want all students to succeed, so we consider student retention and drop out as failures of the system. Adopting an up-or-out system or a widespread program of holding students back would be a major change in American educational policy.

If a test at any level is to be used for student retention, tracking, or remediation, we must assure that students who do not meet or exceed its standards are not just shunted off to dead-end programs with little, if any, hope of returning to the mainstream.

Standard setting in the public school system raises serious questions of educational philosophy and policy. What are high stakes in the fourth grade? What are rewards and sanctions in the fourth grade? If a test at any level is to be used for student retention, tracking, or remediation, we must assure that students who do not meet or exceed its standards are not just shunted off to dead-end programs with little, if any, hope of returning to the mainstream. Using a test for this purpose would necessitate having the resources to use the results for educationally desirable purposes.

At the high school senior level, the idea of a general-purpose exit examination that is used for employment and university admission opposes present legal and professional policies and practices. For an examination to be used in employment, it must be demonstrably valid for that profession. For example, a college entrance examination cannot be used to screen applicants for a janitorial job unless the examination is relevant to required job skills. A very broad-based general exit examination--appropriate for all high school students--would not be likely to identify outstanding students accurately enough for admission to highly selective colleges and universities. Furthermore, if more highly targeted cluster

examinations were used, it would be necessary to demonstrate that each test was a valid predictor of success in college.

Accountability: Questions Requiring Answers

Measuring student attainment of scholastic standards does not easily convert into school accountability.

Using a test for accountability presents a different set of problems. Students at the beginning of their school careers have differences that create as yet unsolved difficulties in using tests for accountability. Some teachers will begin with advantaged and advanced students, and others will not. Are all teachers and schools to be held accountable for having their students reach the same standards? Will expecting equal results be considered fair by teachers and administrators? Will all grades be tested annually so that teachers of certain grades are not judged for the accumulation of student learning over previous grades? How will the migration of some students from school to school within an academic year be handled? Measuring student attainment of scholastic standards does not easily convert into school accountability.

The sometimes conflicting goals of the educational system present dilemmas to administrators. Meeting high academic standards is clearly desirable, but so is preventing students from leaving school before graduation. Since students who drop out are typically low performing, retaining them is likely to increase the probability that a school's students will fail to meet the high standards.

Getting a Grip on the Specifics

It is critical to decide what purposes a test will serve and design a test to meet those goals. A general purpose test intended for many purposes will be very unlikely to serve any purpose well.

A national system that uses high-stakes tests to encourage students to attain high standards is a powerful idea and should be further explored, but many issues must be addressed before such a test can be designed and successfully implemented. Testing is a general technology (Madaus, 1990) that must be tailored to its specific purposes, much like a general-purpose engine that can power automobiles, trucks, or other machines but is more efficient when tailored to a specific use. It is critical to decide what purposes a test will serve and design a test to meet those goals. A general purpose test intended for many purposes will be very unlikely to serve any purpose well.

If a national testing system is to measure student attainment of national and world-class standards, these standards should be specified and a test designed to measure student attainment. Ideally, standards should be set before any new test is constructed.

The educational clout of a test will depend on its stakes, implying important rewards and punishments. We have not yet come to grips with what these stakes will be, especially at the lower grade levels. Given an agreed-upon definition of high-stakes such as promotion, retention, graduation, honors, employment, and university admission, tests can be constructed to serve some or all of these purposes. However, the validity of the test for each purpose would have to be established, since the test would have to be valid and fair at the individual level if individual decisions are to be made. Different tests would be needed for different curricula, and the formal processes of comparing different test results will always be somewhat speculative. Focusing a test on individual decision making will also make monitoring national goals more difficult, if not impossible. Perhaps two types of tests are needed.

Any test should follow from its uses, and a general test for many purposes will not cover any of them well. The potential benefits of tests that drive the curriculum are well worth exploring. As a technology, testing should address the task at hand, but the task at hand is not yet defined well enough to develop instruments to measure it. Focusing on the measuring instrument instead of the attribute measured is not the way to start. Let us begin by addressing the fundamental policy issues, the conflicts and trade-offs in aims and values, and view them within the technical constraints of testing technology. If clear and reasonable test goals are defined, test developers will jump at the opportunity to produce useful instruments, and competition among them will improve testing technology.

References

Beaton, A.E., Hilton, T. L., and Schrader, W. B.(1977). *Changes in the verbal abilities of high school seniors, college entrants and SAT candidates between 1960 and 1972.* Princeton, NJ: Educational Testing Service.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., and York, R. L. (1966). *The equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Linn, R. L. (1991). *Technical considerations in the proposed nationwide assessment system for the national educational goals panel.* (second draft). Mimeo.

Madaus, G. M. (1990). *Testing as a social technology.* Boston, MA: Boston College.

Stocking, M.S., and Mislevy, R. M. (September 20, 1991). Memorandum.